单题项测量：质疑、回应及建议[1]

卫旭华 张亮花
(兰州大学管理学院，兰州 730000)

摘 要 研究者关于单题项测量的争论已久。支持者认为单题项测量具有时间和效率上的优势，而反对者则认为单题项测量的信效度均无法得到保障。通过定性和定量的回顾，归纳了单题项测量的优缺点，剖析了以往研究对单题项测量的种种质疑并逐一进行回应。通过系统梳理，发现单题项测量具有可以接受的信度和效度水平，且多题项测量的效标关联效度并没有显著优于单题项测量。最后，指出了单题项测量开发和使用过程中应该注意的事项。尽管多题项测量仍是当前研究界的主流测量方法，但未来研究者应当更加客观地看待单题项测量。学界应当充分理解单题项测量潜在的优点和适用范围，从而使单题项测量在管理心理学和社会科学研究中发挥其应有的作用。

关键词 单题项；多题项；测量；信度；效度
分类号 B841; B849

构念测量为心理学和社会科学的发展奠定了坚实的方法论基础。在测量构念的时候，研究者通常需要在单题项测量和多题项测量之间做出选择(Ang & Eisend, 2017; Diamantopoulos, Sarstedt, Fuchs, Wilczynski, & Kaiser, 2012; Fuchs & Diamantopoulos, 2009; Sarstedt, Diamantopoulos, & Salzberger, 2016)。单题项测量通常是指仅通过一个测量指标来评价研究构念的方法，具有填答时间短、成本低廉等特点(Bergkvist & Rossiter, 2007)；而多题项测量则是通过两个及以上的题项来评价研究构念的方法，其测量的精度和准确度往往更高(Diamantopoulos et al., 2012; Nair, Ataseven, Habermann, & Dreyfus, 2016)。

虽然两种测量方式都得到了广泛的应用，但二者在实践界和理论界的普及程度却存在较大差异。在应用研究领域，管理咨询者和实务研究者更加偏爱单题项测量，因为它能够大幅降低受访者的拒访率，降低数据收集和处理的成本(Bergkvist & Rossiter, 2007; Hoeppner, Kelly, Urbanoski, & Slaymaker, 2011)。然而，在学术研究领域，研究者更加偏爱多题项测量，有研究指出多数研究构念的题项数介于四到六个之间(Hinkin, 1998)，且伴随着结构方程模型分析技术的普及，多题项测量愈发受到学界的欢迎(Diamantopoulos et al., 2012)。时至今日，有相当部分的研究者认为单题项测量是不全面、不可信的，其信效度都无法得到有效保证(Credé, Harms, Niehorster, & Gaye-Valentine, 2012; Roelen et al., 2014; Roelen et al., 2015)。

由于多题项测量在学术研究领域表现出了压倒性的优势，一些研究者开始尝试比较单题项测量和多题项测量的优劣，以检验传统研究者对单题项测量的质疑是否正确。从二十世纪五十至八十年代的零星比较(Aiken, 1980; Churchill & Peter, 1984; Gorsuch & Mcfarland, 1972; Mosel, 1953)，到近二十年的定性和定量综述(Diamantopoulos et al., 2012; Fuchs & Diamantopoulos, 2009; Wanous, Reichers, & Hudy, 1997)，越来越多的研究证据表明，对于那些界定足够清晰的构念而言，多题项测量并没有显著地优于单题项测量。

自从 Wanous 等(1997)对工作满意度的单题项测量进行了系统回顾之后，单题项测量日益受到了学术研究者的重视和关注，其开发流程和使用方法也日趋完善(Fisher, Matthews, &

[1] 收稿日期: 2018-08-31
* 教育部人文社会科学研究青年基金项目（16YJC630132）；国家自然科学基金青年项目（71602080）；中央高校基本科研业务费专项资金重点项目（17LZUJBWZD007，17LZUJBWZD013）；兰州大学“一带一路”专项项目（2018ldbryb024）资助。
通讯作者: 张亮花, E-mail: zlh331@qq.com

Gibbons, 2016; Fuchs & Diamantopoulos, 2009)。然而，目前国内单题项测量的使用较为混乱，尚无专门讨论单题项测量的研究，这不利于国内研究者了解单题项测量的最新发展动向，也不利于国内研究者正确合理地使用单题项测量。基于此，本文对单题项测量的发展与应用进行了系统回顾，阐述了单题项测量的优缺点，剖析了以往研究的质疑并逐一进行回应，最后指出了单题项测量开发和使用过程中应该注意的事项。

1 单题项测量的优缺点

1.1 单题项测量的优点

尽管多题项测量仍是当前管理心理学研究中的主流测量方法，但在收集数据的过程中，与多题项测量相比，单题项测量能够节约时间和资金成本(Bergkvist & Rossiter, 2007; Franke, Rapp, & Andzulis, 2013; Gogol et al., 2014; Robins, Hendin, & Trzesniewski, 2001)，降低样本偏差并改善数据质量(Postmes, Haslam, & Jans, 2013)。

1.1.1 节约时间

单题项测量在时间上的优势不言而喻，更少的题项会花费答题者更少的时间(Bergkvist & Rossiter, 2007)。这种时间上的优势使得单题项测量具有更高的灵活性(Konstabel, Lönnqvist, Walkowitz, Konstabel, & Verkasalo, 2012)。研究者常常会面对一些特殊的研究问题（如压力、术后的疼痛感等）和研究人群（如老人、儿童、病人、医生等）。这些问题和人群的特殊性可能会对研究者带来时间上的挑战(Fisher et al., 2016; Gogol et al., 2014; Postmes et al., 2013)。比如，调查一个癌症病人的心理压力时，使用多题项测量会增加病人的答题时间，这可能会加重病人的心理压力，进而给病人带来较为负面的影响。此时，使用单题项测量往往能够取得较为理想的效果。

1.1.2 降低资金投入

研究者对答题者的报酬通常视答题者完成题项的数目而定(Wanous et al., 1997)，这是调查研究中常用的规则。在使用多题项测量时，研究者不仅需要加大在问卷印刷过程中的资金投入(Gogol et al., 2014)，还需要支付相对较多的激励费用，从而确保答题者能够认真完成问卷。同时，多题项测量会导致无效问卷数量增加(Postmes et al., 2013)，这些由于答题者的不认真所造成的无效问卷也会增加研究者的资金投入。单题项测量则可以较好地避免上述问题，因此更具经济性。

1.1.3 降低样本偏差

由于答题时间大幅缩短，单题项测量可以有效地降低样本偏差(Postmes et al., 2013)。对于多题项测量而言，很多人不愿意花费时间去完成一份看起来和自己利益无关且很长的问卷。因此，在选择样本的时候，时间因素可能会促使研究者们寻找一些有较多空闲时间的个体，这就造成了样本偏差。与之相反，包含单题项测量的问卷通常较短，这会降低拒访率(Bergkvist & Rossiter, 2007)。单题项测量对答题者的时间要求很低，这样在选择样本的时候，研究者可以选择更加多样化的个体，从而有效地降低样本偏差(Postmes et al., 2013)。

1.1.4 提升数据收集质量

使用单题项测量可以通过节省样本的时间消耗来保证数据收集的有效性。一方面，包含大量多题项测量的问卷会比较冗长，这可能会导致答题者精力无法持续集中，使得他们对靠后的题项需要花费更多时间来阅读和理解(Postmes et al., 2013)，从而产生厌烦感和枯燥感。在这种情形下，答题者可能不会认真地对待问题，而是抱着一种敷衍的态度(Bergkvist, 2015; Bergkvist & Rossiter, 2007; Credé et al., 2012; Gogol et al., 2014; Zimmerman et al., 2006)，并对最终的数据质量产生极大的影响。相反，由于单题项测量花费时间较短，不会使答题者产生较多的厌烦感，因此答题者会更认真地阅读和回答问题(Linden & Rosenthal, 2016)。另一方面，在数据收集完成之后的整理和处理过程中，多题项测量不仅会带来更多的工作量，还会

增加研究者的犯错概率(Reysen, Katzarska-Miller, Nesbit, & Pierce, 2013)，而单题项测量可以为研究者省去许多繁琐的基础工作，从而保证工作的高效性和结果的准确性。

1.2 单题项测量的缺点

由于单题项测量所包含的信息较为有限，它的适用范围受到了限制。批判者指出单题项测量在评价复杂构念和抽象构念时，往往存在较大的局限性(Bergkvist & Rossiter, 2007; Fuchs & Diamantopoulos, 2009)。

1.2.1 不能充分反映复杂构念的内涵

对于一些多维度的复杂构念而言，单一题项很难同时反映这些多维构念不同维度的内涵(Bergkvist & Rossiter, 2007)。单题项测量往往会忽视一些细节问题，比如，当研究者通过一个题项来评价受访者工作满意度的时候，不同的受访者可能会对工作满意度的不同方面进行自动加权处理，给出自己的评价(Fuchs & Diamantopoulos, 2009)。然而，不同受访者的关注点各不相同，对工作满意度各方面的加权处理方式也不尽相同，这就导致研究者无法判断工作满意度的哪个方面更为重要(Oshagbemi, 1999)。比如，同样是工作满意度较低的受访者，一些受访者可能是因为长期得不到晋升而不满，而另外一些受访者可能是因为长期被顾客无理投诉而不满。显然，单题项测量无法识别这些具体的信息。与之对应，多题项测量则能够很好地解决这个问题。

1.2.2 不能准确测量较为抽象的构念

与多题项测量相比，单题项测量仅适用于评价那些足够具体的构念，如薪酬满意度、快乐和焦虑(Gogol et al., 2014; Kim & Abraham, 2017; Wanous et al., 1997)。抽象意味着构念本身的内涵较为模糊，如果通过单题项测量来评价抽象构念，受访者可能会因为过度的模糊性以及理解上的偏差而不能做出正确的判断(Bergkvist & Rossiter, 2007)。研究发现，受访者对抽象构念的理解往往存在较大的异质性(Fuchs & Diamantopoulos, 2009)。对于诸如公司文化、组织公民行为、公共服务动机等抽象构念而言，不同的受访者可能有完全不同的理解，进而导致测量的精度受到影响。

综上，一方面，单题项测量可以降低时间和资金投入，并通过降低答题者的厌烦感来保证数据的质量，从而提高整个研究的效率(Gogol et al., 2014)。另一方面，单题项测量在适用范围上也存在一些不足，它不适合用来评价那些复杂和抽象的构念(Bergkvist & Rossiter, 2007; Fuchs & Diamantopoulos, 2009)。然而，学界对单题项测量最大的质疑在于——很多学者认为单题项测量的信效度无法得到有效保证。在学术研究的评审过程中，使用单题项测量往往被视为是一个重大错误(Wanous et al., 1997)。很多学者在接受研究方法训练时就被告知，应尽量避免使用单题项测量(De Boer et al., 2004)，这些先入为主的观念成为单题项测量备受质疑的重要原因。然而，近些年随着学界对单题项测量数据质量的持续探索，以往关于单题项测量信效度方面的质疑并没有得到实证数据的验证，反而有相当部分的研究发现单题项测量具有可以接受的信效度水平。接下来的部分将着重论述研究者对单题项测量信效度方面的质疑以及回应。

2 对单题项测量信度的质疑及回应

很多研究者认为单题项测量的信度是无法估计的，且单题项测量可能会出现较为严重的测量误差，因此单题项测量被多数研究者拒之门外(Gogol et al., 2014; Van Der Linden & Rosenthal, 2016)。事实上，单题项测量的信度并不是不可估计，随着单题项测量研究的深入，关于单题项测量信度的证据也越来越多，学界需要重新看待单题项测量的信度问题。

2.1 对信度的质疑

信度是衡量一个量表可信程度和稳定性的重要指标(Kim & Abraham, 2017)。研究中最常用的是 Cronbach's α 内部一致性系数，其计算公式如下：

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum\sigma_i^2}{\sigma_x^2}\right]$$

其中，k 为测量题项的数目，σ_i^2 是第 i 个题项分数的方差，σ_x^2 是量表总分的方差。

根据上述公式，计算 Cronbach's α 内部一致性要求题项数必须大于 1，因此，单题项测量的 Cronbach's α 内部一致性是无法计算的。此外，有研究者认为量表的信度和题项数量成正比(Churchill & Peter, 1984)，这会导致单题项测量的信度比多题项测量低(Gogol et al., 2014; Kim & Abraham, 2017; Van Der Linden & Rosenthal, 2016; Wanous et al., 1997)。在变量测量过程中，构念得分的变异包括两个部分：真实得分和随机测量误差。多题项测量可以通过各题项间相互抵消的方式来减少随机测量误差，而单题项测量由于只存在一个题项，无法达到这种效果(Fu, 2005; Konrath, Meier, & Bushman, 2014; Robins et al., 2001)。因此，单题项测量的随机测量误差可能会较为严重，进而导致单题项测量的信度较低。

2.2 对信度的回应

尽管信度是单题项测量常被诟病的一点，但研究者可以通过其它的指标来评价单题项测量的信度，如再测信度或者通过公式来反推单题项测量的信度(Fisher et al., 2016; Spörrle & Bekk, 2014; Wanous et al., 1997)。

2.2.1 再测信度

再测信度体现了量表在时间上的稳定性，通常是通过某个量表在两个时间点得分的相关系数来计算(Fisher et al., 2016)。这种方法是使用最广的衡量单题项测量信度的方法之一。表 1 列出了 60 个单维构念单题项测量的再测信度。从表中可以看到，单维构念单题项测量再测信度高于 0.70 的占比为 68.33%，中位数为 0.73，样本加权平均值为 0.71，说明单题项测量的信度总体上是可信的。

2.2.2 反推信度

研究样本中变量信度的高低会影响研究者对变量间真实关系的估计。在科学研究中，两个变量的真实相关系数可以通过信度修正的样本相关系数求得(Postmes et al., 2013)。真实相关系数与样本相关系数存在如下关系(Loo, 2002; Postmes et al., 2013; Wanous et al., 1997)：

$$\rho = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}$$

其中，r_{xx} 是样本中变量 x 的信度，r_{yy} 是样本中变量 y 的信度，ρ 是 x 和 y 之间的真实相关系数，r_{xy} 是样本中 x 和 y 的相关系数。由于存在测量误差，样本相关系数 r_{xy} 通常小于真实相关系数 ρ。

如果假定 x 和 y 是评价某一构念的两个不同测量，其中，x 是单题项测量，y 是多题项测量，则单题项测量的信度可以通过如下公式来计算：

$$r_{xx} = \frac{{r_{xy}}^2}{\rho^2 \cdot r_{yy}}$$

上式中，由于单题项测量 x 和多题项测量 y 在衡量同一个构念，因此二者的真实相关系数 ρ 的最大值为 1(Postmes et al., 2013; Wanous & Hudy, 2001)。由此可以得到单题项测量信度的最小值：

$$\tilde{r}_{xx} = \frac{{r_{xy}}^2}{r_{yy}}$$

以往关于单题项测量的实证研究中多数都报告了单题项测量与多题项测量的相关系数 r_{xy} 以及多题项测量的信度 r_{yy}，因此，研究者可以通过这种方式来反推单题项测量的最低信度(Loo, 2002; Postmes et al., 2013; Wanous & Hudy, 2001; Wanous et al., 1997)。

通过使用以上方法，本研究对以往实证研究中单题项测量的最小信度进行统计[2]，结果

[2] 用于统计分析的文献已在文后参考文献中标注星号（*）

如表 1 所示。从表中可以看到，仅有 15.38%的单维构念最小信度小于 0.60，其余的均高于 0.60，单维构念单题项测量最小信度的中位数为 0.67。需要指出的是，最小信度并不是单题项测量的实际信度。在实际操作过程中，单题项测量的信度通常高于该最小值。比如，根据 Núñez-Peña, Guilera 和 Suárez-Pellicioni (2014)研究中的数据推断数学焦虑单题项测量的最小信度为 0.63，而其文中报告的再测信度为 0.81，远高于最小信度。然而，对于多维构念，可以看到，其再测信度和最小信度取值普遍偏低。多维构念单题项测量的再测信度和最低信度的中位数分别为 0.62 和 0.50，说明单题项测量在评价多维构念时，可能存在一些问题。由此可见，单题项测量的信度并非像之前学者所提到的那样不可接受，尤其是用来评价单维构念的时候。

表 1 单题项测量的信度统计

信度区间	单维构念				多维构念			
	再测信度		最小信度		再测信度		最小信度	
	数量	百分比	数量	百分比	数量	百分比	数量	百分比
小于 0.60	8	13.33%	2	15.38%	1	33.33%	3	100%
0.60-0.69	11	18.33%	5	38.46%	2	66.67%	—	—
0.70-0.79	26	43.33%	4	30.77%	—	—	—	—
0.80-0.89	11	18.33%	1	7.69%	—	—	—	—
0.90-1.00	4	6.67%	1	7.69%	—	—	—	—
中位数	0.73		0.67		0.62		0.50	

3 对单题项测量效度的质疑及回应

单题项测量不被接受的另外一个原因就是其效度受到质疑。效度是衡量一个测量有效性和准确性的重要指标。单题项测量效度的不足之处主要体现在它的不全面性上(Fisher et al., 2016; Oshagbemi, 1999)。由于单题项测量不能够提供足够多样和细致的答案，一些研究者认为单题项测量忽略了各个被试之间较为细微的差异(Bergkvist & Rossiter, 2007; Diamantopoulos et al., 2012; Konrath et al., 2014; Kwon & Trail, 2005)，这会造成测量结果的偏差，进而使单题项测量的解释力大打折扣。反对者主要从内容效度、构念效度和效标关联效度方面对单题项测量提出质疑。

3.1 对效度的质疑

3.1.1 对内容效度的质疑

内容效度是指测量工具对目标构念内涵与范畴的反映程度。从内容效度来看，单题项测量通常被用于单维构念的测量。尽管单题项测量不够全面，但其依然能够捕获单维构念的核心内涵(Robins et al., 2001)。比如，研究者常常使用单题项测量来评价人们的满意度、生活质量和幸福感等构念(Cheung & Lucas, 2014; De Boer et al., 2004; Zimmerman et al., 2006)。然而，对于那些多维构念，为了尽可能完全地测量多维构念的各个方面，研究者通常使用多个题项来评价多维构念的多个维度。显然，单一题项在反映多维构念的涵义时，其不完全性和片面性尤为显著，内容效度也大打折扣。

3.1.2 对构念效度的质疑

构念效度是考察一个量表是否能够准确测量出目标构念的程度，通常通过聚合效度和区分效度来反映(Fisher et al., 2016)。研究者通常采用组合信度和平均方差萃取量来评价构念测量的聚合效度和区分效度。由于单题项测量无法计算其组合信度和平均方差萃取量(Wohlgemuth & Wenzel, 2016)，因此，不少研究者认为无法评判单题项测量的聚合效度和区分效度(Fuchs & Diamantopoulos, 2009)。

3.1.3 对效标关联效度的质疑

从效标关联效度来看，单题项测量对效标变量所产生的解释力可能更小。有学者指出，单题项测量包含的信息量非常有限(Fisher et al., 2016; Oshagbemi, 1999)，无法全面准确的测量单维或者多维构念，从而影响其对相关变量的预测作用。多题项测量通常包含更多题项和信息，能够更好地反映研究构念的本质，因此具有更好的预测效度(Gogol et al., 2014; Van Der Linden & Rosenthal, 2016)。

正因为上述原因，很多学者对单题项测量的效度表示担忧，从而更青睐选择多题项测量。

3.2 对效度的回应

尽管研究者对单题项测量的效度存在一些质疑，但这些质疑多停留在思辨层面。事实上，近些年已经有相当一部分实证研究发现单题项测量的效度是可以接受的(Elo, Leppänen, & Jahkola, 2003; Konrath et al., 2014; Konstabel et al., 2017; Kwon & Trail, 2005; Robins et al., 2001)。

3.2.1 对内容效度的回应

不可否认，多题项测量可能更加全面，但研究发现多题项测量新增加的题项所产生的额外信息非常小(Fuchs & Diamantopoulos, 2009)。多题项测量通常会设置多个相似但又略有区别的题项来提升测量的全面性。在实际操作过程中，这种相似性往往会造成答题者的厌烦感，答题者常常会在一份问卷中选择相同选项(Nair et al., 2016; Robins et al., 2001)，使得多题项测量获得额外信息的可能性下降。此外，相似题项的设置还存在一个问题，如果这些题项本身是有偏的，多次重复会使这种偏差加倍，造成多题项测量的准确性进一步下降。

近期的实证研究已经证实，单题项测量在评价一些构念时具有可接受的内容效度(Elo et al., 2003; Fisher et al., 2016; Jordan & Turner, 2008; Robins et al., 2001)。比如，Fisher 等(2016)通过专家打分的方法来评价单题项测量的内容效度，结果发现他们新开发的单题项测量拥有良好的内容效度。即便是多维构念，如果其多个维度能够构成一个整体变量，则单题项测量也能够拥有充足的内容效度(Robins et al., 2001)。此外，研究者还通过包含多个形容词的表述方式来增加单题项测量的内容效度。比如，Woods 和 Hampson(2005)通过如下反向计分的单题项测量来评价大五人格中的外向维度："xxx 是一个矜持的、内向的、不喜欢引起人们注意且在陌生人中比较害羞的人"。他们的结果也发现该量表与多题项测量的结果并无太大差异。

3.2.2 对构念效度的回应

为了回应单题项测量构念效度无法评判的质疑，研究者开发出了评判单题项测量构念效度的新方法。首先，在借鉴多特质多方法（MTMM）矩阵的基础上，研究者们通过计算单题项测量和相应的多题项测量的相关系数来衡量单题项测量的聚合效度(Fisher et al., 2016; Postmes et al., 2013)，它反映了单题项测量能够在多大程度上代替多题项测量。如果单题项测量与多题项测量具有较高的相关性，则单题项测量具有较高的聚合效度。其次，研究者通过计算单题项测量与相似构念的相关性来判断单题项测量的区分效度(Fisher et al., 2016; Nichols & Webster, 2013; Spörrle & Bekk, 2014)。如果单题项测量与这些相似构念的相关度较低，则单题项测量具有良好的区分效度。

为了更好地展示单题项测量的构念效度，本研究通过元分析的方法对以往实证研究的聚合效度和区分效度进行了定量整合[3]。表 2 列出了单题项测量的聚合效度与区分效度的定量分析结果。表中的聚合效度代表了单题项测量与多题项测量的相关系数，区分效度代表了单题项测量与相似构念之间的相关性。从表中可以看到，单题项测量与多题项测量高度相关

[3] 表 2 和表 3 中用于元分析的文献已在参考文献中标注星号(*)，这些文献主要是在 Web of Science 和 Google Scholar 中检索获得的，检索时间限定在 1988 年 1 月-2018 年 6 月。效应值编码工作由两位作者独立开展，首轮一致性率超过 90%，不一致的地方由两位作者共同核对修正。元分析软件为 CMA2.0，分析模型为随机效应模型(Borenstein, Hedges, Higgins, & Rothstein, 2011)。

（ρ=0.72，p<0.001），显示出单题项测量具有良好的聚合效度。此外，单题项测量与相似构念之间的相关系数并不高（ρ=0.22，p<0.001），说明单题项测量与这些构念之间具有良好的区分度。因此，以往研究者对单题项测量构念效度的质疑并不成立。

3.2.3 对效标关联效度的回应

即便是多题项测量，其所包含的多个题项很有可能良莠不齐。在对构念进行评估的过程中，好的题项会掩盖不好的题项(Bergkvist & Rossiter, 2007; Fisher et al., 2016; Van Der Linden & Rosenthal, 2016)。然而，单题项测量通常是由专家或者研究者精心挑选出来的题项，通常能够避免不必要的噪音，从而保证构念的预测效度。此外，许多研究者通过比较单题项测量和相应多题项测量在效标关联效度上的差异来评判二者的优劣。已有相当数量的实证研究发现，单题项测量与相应的多题项测量具有相似的效标关联效度(Ang & Eisend, 2017; Gogol et al., 2014; Kwon & Trail, 2005; Riordan et al., 2018; Robins et al., 2001; Woods & Hampson, 2005)。为了更好地展示这一结果，本研究通过元分析方法对以往单题项测量与多题项测量效标关联效度进行定量比较（见表 3）。从表中可以看到，单题项测量和效标变量之间的关系（ρ=0.12，p<0.001）与多题项测量和效标变量之间的关系（ρ=0.16，p<0.001）并无明显差异（Q_B=0.92，p=0.34）。因此，单纯从效标关联效度的角度来看，多题项测量并没有明显优于单题项测量。

表 2 单题项测量的聚合效度与区分效度检验

构念关系	k	N	r	ρ	95%置信区间		Q_W	失安全系数
					低	高		
单题项测量与多题项测量（聚合效度）	45	36930	0.64	0.72***	0.66	0.77	2790.85***	151303
单题项测量与相似构念（区分效度）	51	25225	0.20	0.22***	0.18	0.26	502.70***	12479

注：***表示 p<0.001，k 表示效应值数量，N 为效应值对应的样本量加总，r 为没有经过信度修正的加权相关系数，ρ 为经过信度修正的加权相关系数，Q_W 为组内异质性检验指标，其显著代表不同研究效应值存在显著不同；失安全系数用来评估发表偏差严重程度，当失安全系数大于效应值数量 k 时，代表发表偏差问题并不严重。

表 3 单题项测量与多题项测量的效标关联效度比较

构念关系	k	N	r	ρ	95%置信区间		Q_W	Q_B	失安全系数
					低	高			
单题项测量与效标	63	70478	0.11	0.12***	0.07	0.18	2514.96***	0.92	8323
多题项测量与效标	79	71872	0.14	0.16***	0.11	0.21	3954.72***		20413

注：***表示 p<0.001，Q_B 为组间异质性检验指标，其显著代表单题项测量与效标的关系和多题项测量与效标的关系存在显著差异，其他符号含义见表 2。

4 单题项测量的开发与使用建议

通过前文的回顾，可以看到，尽管单题项测量具有一定的局限性，但是本研究也找到了单题项测量拥有充分信效度的证据。因此，在文献支撑充分和实证证据充足的情况下，研究者可以选取单题项测量来进行他们的研究。

4.1 单题项测量的开发建议

研究者可以从已有的多题项测量中选择一个题项来评价构念(Diamantopoulos et al., 2012; Fisher et al., 2016; Fuchs & Diamantopoulos, 2009; Loo, 2002)，也可以开发一个新的题项来评价构念(Cheung & Lucas, 2014; Gogol et al., 2014; Robins et al., 2001; Van Der Linden & Rosenthal, 2016)。

4.1.1 从已有量表中选择单题项测量

研究者可以从已有表述较为接近或者存在语义重叠的多题项测量中选择因子载荷最高的题项作为单题项测量的评价指标(Diamantopoulos et al., 2012)。这种方法相对较为客观，不会受人为主观因素的干扰。然而，由于因子载荷最高的题项在不同样本中可能会有所不同，研究者可以通过样本加权平均载荷的方法来挑选跨样本最优的题项。此外，研究者也可以通过专家打分的方式从已有的多题项测量中选择内容效度最高的题项(Aiken, 1980; Sarstedt et al., 2016; Sarstedt, Diamantopoulos, Salzberger, Baumgartner, & Woodside, 2016)。研究者可以邀请多位专家，要求他们评价多题项测量中每个题项能够在多大程度上反映构念的核心内涵。随后，研究者可以通过组内一致性指标（rwg）计算专家打分的一致性程度。最后，研究者可以挑选 rwg 得分高于 0.7 且专家打分平均分最高的那个题项作为最终的单题项测量指标。

4.1.2 开发新的单题项测量

研究者也可以通过新开发的单题项测量来评价构念。虽然从多题项测量中获得的题项可能与原来多题项测量的相关性更高，在效标关联效度等指标上也和原来的多题项测量更加接近，但从已有多题项测量中提取单题项测量指标，其描述往往不如新开发的单题项测量指标全面(Fisher et al., 2016)。在产生新的单题项测量的时候，研究者首先要对目标构念有一个正确且清晰地认识，才能够判断它是否适用于单题项测量。研究者应当意识到，一个好的单题项测量应该具有以下特征：能够真实准确地反映构念的核心内涵，能够适用于不同的研究样本，能够跟其他构念有效区分开来(Diamantopoulos et al., 2012)。因此，在描述单题项测量时，研究者要根据构念含义谨慎斟酌用词。在测量一些较为学术化、大众听起来较为生僻的构念时，研究者更应当慎重。否则，单题项测量表述的笼统性以及研究构念的专业性会导致被调查者无法理解研究者所要测量的构念，这就需要研究者对所测量的构念进行额外的书面或者口头解释，反而会浪费更多的时间(Credé et al., 2012)。

需要指出的是，当前实证证据并没有发现单题项测量的量尺会对测量结果产生显著的影响(Kim & Abraham, 2017; Robins et al., 2001)，因此，研究者不必过分担心量尺所带来的研究结果的不确定。

4.2 单题项测量的使用建议

研究者不应一味地抵制单题项测量，应当看到其潜在的优点。与此同时，研究者也应当清醒地认识到，单题项测量并非在每种情况下都适用。

4.2.1 单题项测量更适合单维构念

与多维构念相比，单题项测量更适合那些单一维度的构念。在管理心理学的很多研究中，研究构念通常包含多个不能整合的维度，比如人格、冲突管理风格等。此时，单题项测量只有一个题目，其包含的信息量非常有限，也很难通过一个题项来全面反映多维构念的各个维度。然而，如果多维构念的多个维度能够构成一个整体变量（比如，整体自尊），单题项测量也是可以使用的(Robins et al., 2001)。

4.2.2 单题项测量更适合足够具体且清晰的构念

与过于宽泛、模糊的构念相比，单题项更适合测量足够具体的构念。在社会科学中，有一些研究构念较为抽象，比如“正念”，大众对此类构念的了解非常有限，不适合通过单题项的方式来测量。尽管单题项测量不太适合评价多维构念，但如果多维构念的每个维度足够清晰，研究者可以使用单题项测量来评价多维构念的每一个维度。比如，研究者经常使用单题项测量来评价大五人格中的每一种人格(Konstabel et al., 2017; Konstabel et al., 2012; Woods & Hampson, 2005)，也会使用单题项测量来评价人们的态度（如工作满意度、生活满意度）(Cheung & Lucas, 2014; De Boer et al., 2004; Zimmerman et al., 2006)和情绪（喜怒哀乐）类构念(Fisher & To, 2012; Riordan et al., 2018)。

4.2.3 单题项测量更适合时间受限的研究

当研究者面临一定的资源和时间限制时，可以考虑使用单题项测量对相关构念进行测量。使用单题项测量不仅能够有效地节约时间，还能够很好地解决由于时间问题而引发的样本偏差、资源浪费、不认真作答等问题。单题项测量更加适用于对时间要求较高的实验研究、纵贯研究和经验取样法类研究(Konstabel et al., 2012; Robins et al., 2001)。在不少实验研究中，为了保证实验操控的效果并减少无关因素的干扰，减少实验持续的时间，研究者都采取单题项测量的方法来收集数据。对于纵贯研究和经验取样法类研究，研究者通常需要在一定时期的特定时间点重复收集数据(Fisher & To, 2012)，使用单题项测量能够显著降低答题者的烦躁感和敷衍程度(Fisher et al., 2016)。

4.2.4 单题项测量更适合大样本研究

当前主流的跨国大样本跟踪调查研究有很多（如世界价值观调查）。由于涉及到很多国家的受访者，研究者希望尽可能多的收集数据，导致这些调查会涉及很多研究构念。如果此类调查问卷全部采取多题项测量的方式来衡量，将会给被调查者带来极大的负担。对于大样本调查研究而言，减轻参与者的负担是研究者首要考虑的问题(Cheung & Lucas, 2014)，单题项测量将能够很好地解决这个问题。单题项能够保证关注尽可能多的研究构念，同时又能最大限度地降低被调查者的心理负担。因此，单题项测量往往是大样本研究常用的构念评价方法。

4.2.5 单题项测量更适合研究中的非核心构念

如果一个构念并非研究者关注的重点，或者对其准确程度要求并不高，比如作为一个控制变量，研究者可以优先考虑使用单题项测量(Fisher et al., 2016; Fuchs & Diamantopoulos, 2009)。这样可以让研究者选择尽可能多的控制变量，并在这些已有研究的基础上更好地阐释自己研究变量的独有贡献。

5 结论

通过定性和定量的回顾，本研究总结了单题项测量的优缺点，并对学界关于单题项测量信效度的种种质疑进行了回应。研究发现单题项测量具有可以接受的信度和效度水平，且多题项测量的效标关联效度并没有显著优于单题项测量。尽管多题项测量仍是当前研究界的主流测量方法，但未来研究者应当更加客观地看待单题项测量。研究者应当看到单题项测量潜在的优点和适用范围，从而使单题项测量在管理心理学和社会科学研究中发挥其应有的作用。

致谢：感谢香港树仁大学工商管理系 Kelly Z. Peng 副教授对本文英文摘要的修订！

参考文献

*表示定量分析纳入的文献

*Abdelkhalek, A. M. (2006). Measuring happiness with a single-item scale. *Social Behavior and Personality: An International Journal, 34*(2), 139–150.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement, 40*(4), 955–959.

Ang, L., & Eisend, M. (2017). Single versus multiple measurement of attitudes: A meta-analysis of advertising studies validates the single-item measure approach. *Journal of Advertising Research, 58*(3), 1–10.

Bergkvist, L. (2015). Appropriate use of single-item measures is here to stay. *Marketing Letters, 26*(3), 245–255.

Bergkvist, L., & Rossiter, J. R. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*(2), 175–184.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to meta-analysis*. United Kingdom: John Wiley & Sons.

*Cheung, F., & Lucas, R. E. (2014). Assessing the validity of single-item life satisfaction measures: Results from three large samples. *Quality of Life Research, 23*(10), 2809–2818.

Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research, 21*(4), 360–375.

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the big five personality traits. *Journal of Personality and Social Psychology, 102*(4), 874–888.

*De Boer, A. G. E. M., Van Lanschot, J. J. B., Stalmeier, P. F. M., Van Sandick, J. W., Hulscher, J. B., De Denissen, J. J. A., Geenen, R., Selfhout, M., & Aken, M. A. G. V. (2008). Single-item big five ratings in a social network design. *European Journal of Personality, 22*(1), 37–54.

De Boer, A. G., Van Lanschot, J. J., Stalmeier, P. F., Van Sandick, J. W., Hulscher, J. B., de Haes, J. C., et al. (2004). Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research, 13*(2), 311–320.

*Denissen, J. J. A. , Geenen, R. , Selfhout, M. , & Aken, M. A. G. V. . (2010). Single-item big five ratings in a social network design. *European Journal of Personality, 22*(1), 37–54.

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science, 40*(3), 434–449.

Elo, A. L., Leppänen, A., & Jahkola, A. (2003). Validity of a single-item measure of stress symptoms. *Scandinavian Journal of Work Environment & Health, 29*(6), 444–451.

Fisher, C. D., & To, M. L. (2012). Using experience sampling methodology in organizational behavior. *Journal of Organizational Behavior, 33*(7), 865–877.

*Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23.

Franke, G. R., Rapp, A., & Andzulis, J. M. (2013). Using shortened scales in sales research: Risks, benefits, and strategies. *Journal of Personal Selling & Sales Management, 33*(3), 319–328.

Fu, Y. (2005). Measuring personal networks with daily contacts: A single-item survey question and the contact diary. *Social Networks, 27*(3), 169–186.

Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft, 69*(2), 195–210.

*Gardner, D. G., Cummings, L. L., Dunham, R. B., & Pierce, J. L. (1998). Single-item versus multiple-item measurement scales: an empirical comparison. *Educational and Psychological Measurement, 58*(6), 898–915.

*Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., et al. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology, 39*(3), 188–205.

Gorsuch, R. L., & Mcfarland, S. G. (1972). Single vs. Multiple-item scales for measuring religious values. *Journal for the Scientific Study of Religion, 11*(1), 53–64.

*Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*(6), 504–528.

*Haes, J. C. J. M., & Sprangers, M. A. G. (2004). Is a single-item visual analogue scale as valid, reliable and responsive as multi-item scales in measuring quality of life? *Quality of Life Research*, *13*(2), 311–320.

Hinkin, T. R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods, 1*(1), 104–121.

*Hoeppner, B. B., Kelly, J. F., Urbanoski, K. A., & Slaymaker, V. (2011). Comparative utility of a single-item vs. Multiple-item measure of self-efficacy in predicting relapse among young adults. *Journal of Substance Abuse Treatment, 41*(3), 305–312.

*Jordan, J. S., & Turner, B. A. (2008). The feasibility of single-item measures for organizational justice. *Measurement in Physical Education and Exercise Science, 12*(4), 237–257.

*Kim, H. J., & Abraham, I. (2017). Measurement of fatigue: Comparison of the reliability and validity of single-item and short measures to a comprehensive measure. *International Journal of Nursing Studies, 65*, 35–43.

*Konrath, S., Meier, B. P., & Bushman, B. J. (2018). Development and validation of the single item trait empathy scale (SITES). *Journal of Research in Personality,* 73, 111–122.

*Konrath, S., Meier, B. P., & Bushman, B. J. (2014). Development and validation of the single item narcissism scale (SINS). *Plos One, 9*(8), e103459.

Konstabel, K., Lönnqvist, J. E., Leikas, S., García, V. R., Qin, H., Verkasalo, M., et al. (2017). Measuring single constructs by single items: Constructing an even shorter version of the "short five" personality inventory. *Plos One, 12*(8), e0182714.

Konstabel, K., Lönnqvist, J. E., Walkowitz, G., Konstabel, K., & Verkasalo, M. (2012). The 'short five' (S5): Measuring personality traits using comprehensive single items. *European Journal of Personality, 26*(1), 13–29.

*Kwon, H., & Trail, G. (2005). The feasibility of single-item measures in sport loyalty research. *Sport Management Review, 8*(1), 69–88.

Linden, S. V. D., & Rosenthal, S. A. (2016). Measuring narcissism with a single question? A replication and extension of the single-item narcissism scale (SINS). *Personality and Individual Differences, 90*(3), 238–241.

*Loo, R. (2002). A caveat on using single-item versus multiple-item scales. *Journal of Managerial Psychology, 17*(1), 68–75.

*Macias, C., Gold, P. B., Öngür, D., Cohen, B. M., & Panch, T. (2015). Are single-item global ratings useful for assessing health status? *Journal of Clinical Psychology in Medical Settings, 22*(4), 1–14.

*Matthias, S., & Magdalena, B. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment, 21*(3), 272–285.

Mosel, J. N. (1953). Single-item tests for personnel screening. *Educational and Psychological Measurement, 13*(2), 179–192.

*Núñez-Peña, M. I., Guilera, G., & Suárez-Pellicioni, M. (2014). The single-item math anxiety scale (SIMA): An alternative way of measuring mathematical anxiety. *Personality and Individual Differences, 60*(4), S75–S76.

*Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77–86.

Nair, A., Ataseven, C., Habermann, M., & Dreyfus, D. (2016). Toward a continuum of measurement scales in just-in-time (JIT) research – an examination of the predictive validity of single-item

and multiple-item measures. *Operations Management Research, 9*(1–2), 35–48.

*Nichols, A. L., & Webster, G. D. (2013). The single-item need to belong scale. *Personality and Individual Differences, 55*(2), 189–192.

Oshagbemi, T. (1999). Overall job satisfaction: How good are single versus multiple-item measures? *Journal of Managerial Psychology, 14*(5), 388–403.

*Postmes, T., Haslam, S. A., & Jans, L. (2013). A single-item measure of social identification: Reliability, validity, and utility. *British Journal of Social Psychology, 52*(4), 597–617.

*Reysen, S., Katzarska-Miller, I., Nesbit, S. M., & Pierce, L. (2013). Further validation of a single-item measure of social identification. *European Journal of Social Psychology, 43*(6), 463–470.

Riordan, B. C., Cody, L., Flett, J. A. M., Conner, T. S., Hunter, J., & Scarf, D. (2018). The development of a single item FoMO (Fear of Missing Out) scale. *Current Psychology,* doi: 10.1007/s12144-018-9824-8.

*Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin, 27*(2), 151–161.

Roelen, C. A., Van, R. W., Groothoff, J. W., Jj, V. D. K., Twisk, J. W., & Heymans, M. W. (2014). Work ability as prognostic risk marker of disability pension: Single-item work ability score versus multi-item work ability index. *Scandinavian Journal of Work Environment & Health, 40*(4), 428–431.

Roelen, C. A. M., Heymans, M. W., Twisk, J. W. R., Laaksonen, M., Pallesen, S. L., Magerøy, N., et al. (2015). Health measures in prediction models for high sickness absence: Single-item self-rated health versus multi-item SF-12. *European Journal of Public Health, 25*(4), 668–672.

*Russell, J. A. (1989). Affect grid: a single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology, 57*(3), 493–502.

Sarstedt, M., Diamantopoulos, A., & Salzberger, T. (2016). Should we use single items? Better not. *Journal of Business Research, 69*(8), 3199–3203.

Sarstedt, M., Diamantopoulos, A., Salzberger, T., Baumgartner, P., & Woodside, A. G. (2016). Selecting single items to measure doubly concrete constructs: A cautionary tale. *Journal of Business Research, 69*(8), 3159–3167.

Spörrle, M., & Bekk, M. (2014). Meta-analytic guidelines for evaluating single-item reliabilities of personality instruments. *Assessment, 21*(3), 272–285.

*Van Der Linden, S., & Rosenthal, S. A. (2016). Measuring narcissism with a single question? A replication and extension of the single-item narcissism scale (SINS). *Personality and Individual Differences, 90*(3), 238–241.

*Wanous, J. P., & Hudy, M. J. (2001). Single-item reliability: A replication and extension. *Organizational Research Methods, 4*(4), 361–375.

*Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single-item measures? *Journal of Applied Psychology, 82*(2), 247–252.

Wohlgemuth, V., & Wenzel, M. (2016). Dynamic capabilities and routinization. *Journal of Business Research, 69*(5), 1944–1948.

*Woods, S. A., & Hampson, S. E. (2005). Measuring the big five with single items using a bipolar response scale. *European Journal of Personality, 19*(5), 373–390.

Zimmerman, M., Ruggero, C. J., Chelminski, I., Young, D., Posternak, M. A., Friedman, M., et al.

(2006). Developing brief scales for use in clinical practice: The reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. *Journal of Clinical Psychiatry, 67*(10), 1536–1541.

附录：定量分析文献查找、编码及分析过程

1、文献查找过程

研究者在 Web of Science、谷歌学术、EBSCO、Springer、Sage、Elsevier、CNKI 和百度学术等数据库检索已发表的单题项测量的文献，检索的关键词为 single-item measure（单题项测量）或 single-indicator measure（单指标测量），时间限定在 1988 年 1 月-2018 年 6 月之间发表的文献。此外，我们还通过阅读一些单题项测量的综述类文章进行查漏补缺，以确保包含尽可能多的单题项测量文章。在检索到的文章中，只要满足如下任一标准，均纳入了我们的定量分析：（1）报告了单题项测量的再测信度；（2）报告了单题项测量与相同构念多题项测量的相关系数；（3）报告了单题项测量与相似构念的相关系数；（4）报告了单题项测量以及多题项测量与一些效标变量的相关系数。根据这些标准，我们一共筛选出了 29 篇符合要求的单题项测量定量研究论文。纳入分析的文章一共报告了 79 个与信度相关的数值，45 个检验聚合效度的效应值，51 个检验区分效度的效应值，63 个单题项测量的效标关联效度效应值以及 79 个多题项测量的效标关联效度效应值。

2、变量编码过程

两位研究者首先依据上述编码标准独立开展效应值编码工作，首轮一致性率超过 90%。对于编码过程中出现的不一致，两位作者共同核对后进行修正，最终形成了一致的编码结果。在信度编码过程中，单题项测量的再测信度主要用于再测信度的统计。此外，如果文章既报告了单题项测量与相同构念多题项测量的相关系数，又报告了相同构念多题项测量的信度系数，我们可以依据这些数据计算单题项测量反推的最小信度（计算公式详见正文 2.2.2 节）。在效度编码过程中，如果作者分别用单题项测量和多题项测量来评价同一构念，且文中报告了二者的相关系数，我们可以依据此数据来计算聚合效度的加权平均效应值。如果文章报告了单题项测量与相似构念的相关系数，我们可以依据此数据来计算区分效度的加权平均效应值。如果文章报告了单题项测量以及多题项测量与一些效标变量的相关系数，我们可以据此统计效标关联效度的加权平均效应值，并对单题项测量和多题项测量效标关联效度的差异进行检验。当同一个样本同一个构念出现多个相似测量时，本研究将这些相似测量的效应值进行合并，以减少人为扩大样本量而对研究结果造成的干扰。此外，我们还收集了多题项测量的信度，并对效应值进行了基于测量误差的修正。对于那些未报告信度值的多题项测量，我们以其他多题项测量的平均加权信度（0.762）来替代其信度水平。

3、分析过程

本研究单题项测量的信度分析为频数统计，结果详见正文表 1。单题项测量的效度分析为元分析，分析软件为 CMA2.0，元分析结果详见正文表 2 与表 3。在元分析的过程中，我们通过失安全系数来检验本研究所涵盖论文发表偏差问题的严重程度。正文表 2 和表 3 结果显示，每组分析的失安全系数均大于其对应的效应值数量，说明本研究的发表偏差问题并不严重。在元分析模型的选择上，根据 Borenstein, Hedges, Higgins 和 Rothstein(2011)的建议，我们选取了随机效应模型。该模型假定不同研究中的效应值是随机变化的，而非固定不变的，与现实情况的吻合度更高。

Single-item measures: Queries, responses and suggestions

WEI Xuhua; ZHANG Lianghua

(*School of Management*, *Lanzhou University*, *Lanzhou 730000*, *China*)

Abstract: Single-item measures have long been debated by researchers. Proponents advocate the efficiency of single-item measures, while opponents question reliability and validity of them. Through comprehensive qualitative and quantitative reviews, advantages and disadvantages of single-item measures are summarized, and doubts and debates in the literature are analyzed and responded specially. Through systematic review, results show that single-item measures have acceptable level of reliability and validity. More notably, the criterion-related validity of multi-item measures is not superior than single-item measures. With the review and current findings, issues that should be noted during the development and usage of single-item measures were addressed. Although multi-item measures are more acceptable, single-item measures may also have its legitimacy in academic research. That is, researchers should value the potential advantages and application scope of single-item measures, so that single-item measures can play its appropriate role in managerial psychology and social science research.

Keywords: single-item; multi-item; measure; reliability; validity